Empirically testing *Tonnetz*, voice-leading, and spectral models of perceived triadic distance

Submitted exclusively to the *Journal of Mathematics and Music*
Last compiled on February 5, 2016

Empirically testing *Tonnetz*, voice-leading, and spectral models of perceived triadic distance

Andrew J. Milne[a*] and Simon Holland[b]

[a]*MARCS Institute, University of Western Sydney, Sydney, Australia;*
[b]*Music Computing Lab, Centre For Research In Computing, The Open University, Milton Keynes, UK*

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

We compare three contrasting models of the perceived distance between root-position major and minor chords and test them against new empirical data. The models include a recent psychoacoustic model called spectral pitch-class distance, and two well-established music theoretical models – *Tonnetz* distance and voice-leading distance. To allow a principled challenge, in the context of these data, of the assumptions behind each of the models, we compare them with a simple "benchmark" model that simply counts the number of common tones between chords. Spectral pitch-class distance and *Tonnetz* have the highest correlations with the experimental data and each other, and perform significantly better than the benchmark. The voice-leading model performs worse than the benchmark. We suggest that spectral pitch-class distance provides a psychoacoustic explanation for perceived triadic distance and its music theory representation, the *Tonnetz*. The experimental data and the computational models are available in the online supplement.

Keywords: spectral pitch-class distance; voice-leading distance; *Tonnetz*; similarity; fit; harmony

2010 Mathematics Subject Classification: 15A04; 15A06; 65C20
2012 Computing Classification Scheme: Applied computing~Sound and music computing

1. Introduction

Since Aurelian's designation of pitches as "high" and "low" in the middle of the ninth century (Cohen 2002), distance-based representations of musical pitch have played an important role in music theory, pedagogy, notation, models of music perception, and musical instrument design. For instance, pitches have typically been represented as points on an idealized line extending from low to high, or on a curve such as a helix to additionally model the similarity of octaves (e.g. Drobisch 1855, Shepard 1982, and Deutsch 1982), or on a circle (a flattened helix) in contexts where pitches an octave apart can be usefully considered equivalent (e.g. Révész 1913 and Bachem 1950). In many cases, the structure is chosen so that the distances between pitches approximate their perceived musical distances. The importance of such representations is that they can facilitate the understanding of music for musicians and composers; they can also shed light on processes that may underlie our perception and cognition of music.

Distance relationships can also be applied to chords, where they can play a similarly

*Corresponding author. Email: a.milne@westernsydney.edu.au

useful role. Major and minor triads, in particular, are often described as one of the fundamental units in the construction and appreciation of tonal music, but their relationships are arguably more complex than those of individual pitches. The line- and curve-based models for perceived pitch distance seem intuitively effective and they are relatively simple. However, for chords, there are a variety of different possible models. If one were to consider successive chords solely as multiple voices (melodic lines), voice-leading models (which total the pitch distances moved by all voices) would be an obvious choice. However, a chord – notably a major or minor triad – is not just the sum of its parts. It has a *root*, which is a specific pitch class that is the most salient and determines the perceived stability of the chord depending on whether or not it is the lowest pitch (Parncutt 1988); these different weightings and chord stabilities may play an additional role. Furthermore, Huron (1989) and McLachlan, Marco, and Wilson (2012) have demonstrated that even musically experienced listeners cannot reliably enumerate the pitches in an unfamiliar chord, let alone individuate their pitches (indeed, this is an aural skill that often requires considerable training). These two aspects of harmony suggest that pure voice-leading models may not be the most appropriate method for modelling perceived triadic distances. Analogous complications also hold for the perceived distances between keys.

A common alternative, for both chord and key distances, is the chord *Tonnetz* (German for *tone network*). The chord *Tonnetz* is a regular geometrical lattice where chords are represented by points such that chords with roots a perfect fifth/fourth, or third/sixth apart, and those sharing two common tones (e.g. C maj and A min), are closer than those without. The historical longevity of, and current interest in, voice-leading and *Tonnetz*-like models in the domain of music theory (e.g. Cohn 1997; Tymoczko 2006; Callender, Quinn, and Tymoczko 2008; Tymoczko 2011; Capuzzo 2014) suggest they are worthy of investigation in an experimental context.

In this paper, we also put forward a psychoacoustically oriented *spectral* model, which treats each chord (or tone) as a large collection of pitch classes corresponding to all its tones' harmonics (frequency components). Other recent research has demonstrated that spectral models can successfully predict the perceived fit of chromatic probe tones to an established key (Krumhansl's (1982) tonal hierarchies) (Milne, Laney, and Sharp 2015) and the perceived affinity of microtonal melodies (Milne, Laney, and Sharp 2016). The spectral model tends to make chords with more common tones and more tones a perfect fifth/perfect fourth apart (between any voices) closer than those with fewer such intervals. (Our precise formalizations of all these models are detailed subsequently.)

1.1. *Symmetry and asymmetry of distance*

According to most definitions, distance is symmetrical with respect to order. This means that the distance from object x to object y is equal to the distance from y to x. Familiar spatial distances (e.g. Euclidean) are also invariant under translation (in other words, the distance between x and y is equivalent to the distance between $x + a$ and $y + a$). (Both properties hold for all the models we test in this paper.) If true for perceived distances between chords, these properties would mean that neither the order nor the transposition of the chord pair would matter. This raises an immediate question: is a symmetrical, translation invariant measure like this directly applicable to music perception in which asymmetrical relationships play a vital role?[1]

For example, Bharucha and Krumhansl (1983) have shown that, given a previously

[1] A question also raised by Krumhansl (1990).

established context such as a C major key (established with a IV-V-I cadence), the chords G maj and C maj are "closer" than the chords D maj and G maj. Furthermore, in a C major context, conventional tonal intuition suggests that the chord C maj will be heard to follow G maj better than G maj follows C maj (the former being the stable authentic cadence in Western music, the latter being a less stable half-cadence). These suggest that asymmetry plays an important role in music perception. However, as argued below, asymmetrical features of music perception can arise from combinations of symmetrical distances.

Simply put, apparent asymmetries in the distance between two objects may occur due to their both being measured with respect to a third object. For example, suppose the perceived distance between the chords C maj and G maj is the same as the perceived distance between G maj and D maj (this is to be expected because the latter pair is just a transposition – mathematically, a translation – of the former pair). But now suppose that both chord pairs are played directly after a previously established C major key context, which we represent by the pitch-class set C, D, E, F, G, A, B. It is likely the D maj chord is more distant from this context than the C maj chord because, for example, only the former contains a pitch class (F♯) not in the scale; this may well result in the distance between G maj and D maj being rated as larger than the distance between the C maj and G maj. Even though both the chord-chord and chord-scale distances are individually invariant with respect to translation, their combination is not.

In the second of the examples noted above (namely, that in a C major key context, the chord C maj will likely be heard to follow G maj better than G maj follows C maj) the apparent order asymmetry may occur because, in relation to a C major scale context (used, as before, to represent the C major key), the first pair moves from a possibly more distant chord-scale relationship (G maj chord in a C major scale) to a possibly closer chord-scale relationship (C maj chord in a C major scale), while the second pair moves in the opposite direction and hence from a closer relationship to a more distant relationship.[2] Analogously to the previous example, both the chord-chord and chord-scale distances are individually symmetrical with respect to order, but their combination is not.

It is possible that, in harmony perception, other embedded contexts may play similar roles; for example, each pitch class is contextualized by the chord of which it is a part, the chords that precede and follow it, the scale from which each of these chords is taken, the broader and most dominant key or scale of the entire piece of music, and the prevalent cultural practice of which the music is an exemplar (in addition, episodic memories, the acoustical environment, and the listener's neurological make-up, as well as numerous other factors, may provide important additional contexts). Each of these contexts may introduce surface level asymmetries, which are actually combinations of symmetrical distances between multiple deeper levels.

For these reasons, we believe that investigating the more atomic symmetrical distances between musical objects (in this case, root-position triads) is a vital task. This is because combining such distances across different musical levels may allow us to build more complex models that elucidate the complex and apparently asymmetrical perception of harmonic relationships in tonal (and microtonal) music. An example of such a methodology is given in Milne, Laney, and Sharp (2015) with their model of the tonicness of chords in a variety of conventional and microtonal scales.

[2] Although not relevant to the more abstract argument being made here, the spectral pitch-class model does indeed predict that the C maj chord is closer to the C major scale than is the G maj chord (Milne, Laney, and Sharp 2015).

1.2. *Experimental methods and related research*

The overall purpose of the study is to compare three principal kinds of models of musical distance and to test them against experimentally obtained ratings of the perceived distance between all root-position major and minor triads. As noted earlier, the models (detailed in the next section) include a recent psychoacoustic model called spectral pitch-class distance, and two well-established music theoretical models: *Tonnetz* distance and voice-leading distance.

We do not seek to find a single best model for all aspects of perceived triadic distance. Such a thing is somewhat nebulous in that its meaning as well as its "quantity" will vary depending on the listening context, the aim of the listening (analytical, compositional, for pleasure, etc.), as well as the personality and cultural context of the listener. In this paper, however, we do aim to find the best model for the data we have collected. As detailed later, listeners were presented with naturalistic sounding material and were asked to listen attentively. We expect, therefore, that the better a model performs on these data, the more likely it is that it will also adequately represent and explain aspects of real-world listeners' responses to chord progressions in music with which they are actively engaging. Furthermore, the models are available for download and so can also be tested on new empirical data.

We have endeavoured to construct a set of experimental methods (fully described in Sec. 3) that allow us to measure precisely the posited underlying symmetrical aspects of harmony perception described in the previous section. We have also attempted to minimize the possibility that the principle embodied by any single model may gain an unfair advantage. This means that in many respects our methods differ from those used in previous related research by Krumhansl and Kessler (1982), Bharucha and Krumhansl (1983), Bigand, Parncutt, and Lerdahl (1996), Krumhansl (1998), and Rogers and Callender (2006). Notably, (as motivated and detailed in Sec. 3.3) we used stimuli with full harmonic complex tones rather than octave complex tones, we repeatedly looped the chords, we did not establish a tonal context (indeed, we took measures to avoid such a context inadvertently arising), and we asked our participants to make ratings about the relationship between the chords (e.g. their similarity) rather than a property of one of the chords (e.g. its tension). We are unaware of any previous research that has done all of these things together and explicitly attempted to measure the perceived symmetrical distance between major and minor triads.

2. The models

When comparing the three principal kinds of model of perceived triadic distance (spectral, *Tonnetz*, and voice-leading) with our participants' ratings, each model was first parameterized to ensure that all well-known variants of each model could be tested against the data, and the best fitting variant of each model selected. Potential over-flexibility was shown to be negligible by a process of cross-validation (in Sec. 4.3).

In addition to the parameterizations of all of the models, the voice-leading class of model was further split into two variant models, representing different assumptions about how voice-leading may operate. A variant of the *Tonnetz* model based on transformational distance was also created. A sixth and final model, the Hamming model, was included specifically as a benchmark model. The Hamming model fits this role well, since it contains only elements common to all of the other models – each of the other models adds additional assumptions to this benchmark. Consequently, the failure of any

model to outperform the benchmark would suggest its assumptions are inapplicable to this context.

Thus, the six models and their parameterizations are detailed in the following subsections. The models, empirical data, and statistical analyses have been coded in MATLAB and can be downloaded from the online supplement or http://www.dynamictonality.com/harmonic_distance_files/.

2.1. *The* Tonnetz *and its parameterization*

The *Tonnetz* comes in two standard forms: the *pitch-class* Tonnetz, whose points represent pitch classes, and the *chord* Tonnetz, whose points represent major or minor triads. It is the latter that is our principal interest in this paper, though we need to understand the former because the latter is most easily understood as being derived from it.

The *pitch-class* Tonnetz is a geometric structure which has many closely related versions. In all cases, pitch classes are points in a two-dimensional lattice with axes of a perfect fifth and a major (or minor) third. However, the fifth and third axes are not necessarily geometrically orthogonal. For example, the canonical form is perfectly hexagonal, with the fifth and third axes at 60 degrees but, historically, a variety of other instantiations have been used (e.g. Euler 1739, Oettingen and Riemann (see Gollin 2011), Longuet-Higgins 1962, Balzano 1980, Holland 1994, Chew 2006, etc.). The *chord* Tonnetz can be derived from the pitch-class *Tonnetz* in that each chord lies at the centroid of the shape enclosed by its pitch classes. The *centroid* can be characterized as the mean position of all points in a shape; alternatively, it can be thought of as the centre of mass (assuming the shape has uniform density and thickness).[3]

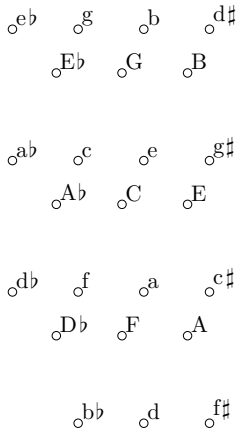
In order to allow our *Tonnetz* model to be parameterized to encompass many typical variations, we start with the canonical form but then allow for the perfect fifth axis to be independently scaled (by parameter s), and for the lattice to be sheared parallel with the perfect fifth axis (by parameter h). The effects of a small but indicative sample of these transformations on the resulting chord *Tonnetz* are illustrated in Figure 1. The middle row has an identity scaling of 1, the top row has a scaling of $\sqrt{3} \approx 1.73$, the bottom row has a scaling of $\sqrt{1/3} \approx 0.58$; the central column has an identity shear of 0, the left column has a shear of $-\sqrt{1/3} \approx -0.58$, the right column has a shear of $\sqrt{1/3} \approx 0.58$. Uppercase letters are major triads, lowercase letters are minor triads.

The *Tonnetz distance* between any two chords is simply defined as their Euclidean distance (i.e. the straight-line, as-the-crow-flies, distance). (After the formalization of the Euclidean approach, we will discuss an alternative *Tonnetz* distance measure.) However, a complicating factor is that any given pitch class will appear at more than one location in an extended (or infinite) *Tonnetz*. For example, in Figure 1(e), note how the chords D and f♯, appear twice, as also do numerous enharmonically equivalent chords, such as G♭/F♯, and g♯/a♭.

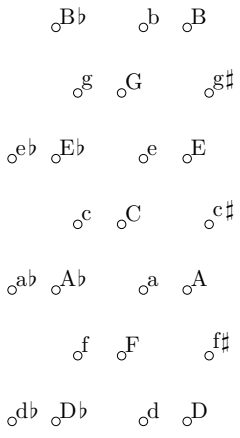
For many uses of a *Tonnetz*, this duplication can be ignored trivially by choosing a subset of the plane that contains just one member of each enharmonic equivalence class.[4] However, crucially, when measuring the *Tonnetz* distance between any two pitch classes

[3]The two types of *Tonnetz* can also be characterized as geometrical duals (Cohn 1998; Tymoczko 2012).

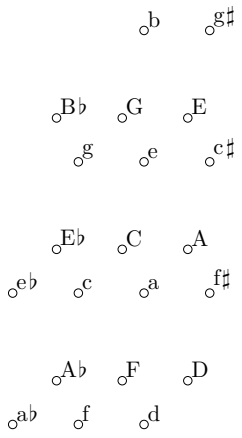
[4]In this paper, *enharmonic equivalence* refers to all *Tonnetz* locations with the same pitch in 12-tone equal temperament, which is the tuning of the experimental stimuli (in just intonation, their pitches would differ). This includes standard enharmonic equivalences such as ..., B♯♯♯, C♯♯, D, E♭♭, F♭♭♭, ..., but it also includes *Tonnetz* locations with the same pitch name. More precisely, all notes separated by some integer combination of the *syntonic comma* (up four *Tonnetz* fifths and down one *Tonnetz* major third) and the *major diesis* (up four *Tonnetz* fifths and down four *Tonnetz* major thirds) are termed enharmonically equivalent because they have the same pitch in 12-TET. They have the same pitch because these two commas are a basis of the null space of the



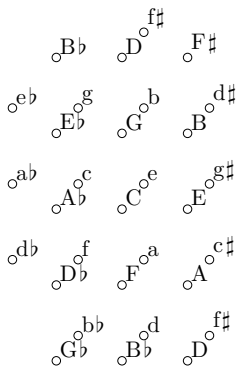
(a) Shear $= -\sqrt{1/3}$, scale $= \sqrt{3}$.



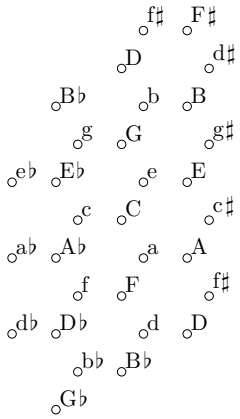
(b) Shear $= 0$, scale $= \sqrt{3}$. The Balzano chord Tonnetz.



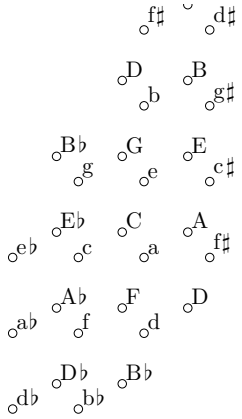
(c) Shear $= \sqrt{1/3}$, scale $= \sqrt{3}$.



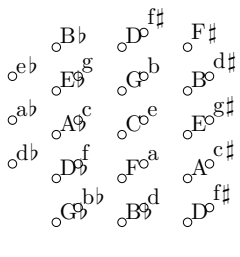
(d) Shear $= -\sqrt{1/3}$, scale $= 1$. Similar to Oettingen's pitch-class Tonnetz (Gollin 2011).



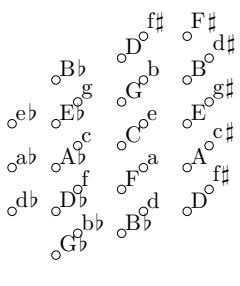
(e) Shear $= 0$, scale $= 1$. The canonical hexagonal chord Tonnetz (e.g. Douthett and Steinbach (1998); Tymoczko (2012)). Similar to Riemann's pitch-class Tonnetz (Gollin 2011).



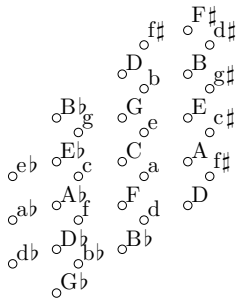
(f) Shear $= \sqrt{1/3}$, scale $= 1$.



(g) Shear $= -\sqrt{1/3}$, scale $= 1/\sqrt{3}$. Similar to Euler's (1739) pitch-class Tonnetz.



(h) Shear $= 0$, scale $= \sqrt{1/3}$. Similar to the chord Tonnetz optimized to our empirical data (see Fig. 6.)



(i) Shear $= \sqrt{1/3}$, scale $= \sqrt{1/3}$.

Figure 1. Some examples showing the effects of the shear and scale parameters on the chord *Tonnetz*. Uppercase letters are major triads, lowercase are minor triads. The parameters' effects on the pitch-class *Tonnetz* are identical, and can be visualized by ignoring the minor triads.

or chords, we should choose the shortest amongst all possible enharmonic equivalents. Moreover, as the scale and shear parameters are varied, the location of that shortest path may change. Consequently, in our distance definition, we need to explicitly consider the issue of enharmonically equivalent pitch classes in the *Tonnetz*. Thus, in order to find the variant of the *Tonnetz* model that best fits the empirical data (i.e. finding its optimal parameter values), the *Tonnetz* distances need to be calculated between a wide variety of enharmonic equivalents for every different setting of the scale and shear parameters. On a technical note, the resulting nonlinearity of distance as a function of the scale and shear parameters means that the model must be optimized – to empirical data – iteratively (e.g. using a gradient descent method) rather than analytically.

The above considerations are concisely formalized in the following equation. Directly after the equation, we define the variables and then provide a brief walk-through of how the formula works:

$$d_{\mathrm{T}}(\boldsymbol{x}, \boldsymbol{y}; h, s) = \min_{\boldsymbol{q} \in \mathcal{Q}} \left(\|\mathbf{HSL}(\boldsymbol{q} + g\boldsymbol{c})\|_2 \right) ; \tag{1}$$

- $d_{\mathrm{T}}(\boldsymbol{x}, \boldsymbol{y}; h, s)$ is the modelled *Tonnetz* distance between the chords $\boldsymbol{x}$ and $\boldsymbol{y}$, as parameterized by the shear $h \in (-\infty, \infty)$ and scale $s \in [0, \infty)$ values;
- $\mathbf{H} = \left(\begin{smallmatrix} 1 & 0 \\ h & 1 \end{smallmatrix}\right)$ is a shear matrix containing the shear parameter h, such that when $h = 0$ no change occurs;
- $\mathbf{S} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & s \end{smallmatrix}\right)$ is a scale matrix containing the scale parameter s, such that when $s = 1$ no change occurs;
- $\mathbf{L} = \left(\begin{smallmatrix} 0 & \sqrt{3/4} \\ 1 & 1/2 \end{smallmatrix}\right)$ transforms coordinates from a square *Tonnetz* (with perfect fifths and major thirds as axes) into corresponding coordinates from the canonical hexagonal *Tonnetz* (which has unit distances between pitch classes separated by fifths or thirds, and the perfect fifth and major third axes are at 60°);
- $\boldsymbol{q} \in \mathbb{Z}^2$ gives the numbers of perfect fifths and major thirds required to get from the root of chord $\boldsymbol{x}$ to the root of chord $\boldsymbol{y}$ (i.e. their displacement in a square *Tonnetz*). Since each pitch class occurs at different locations on the *Tonnetz*, $\boldsymbol{q}$ has more than one possible value for any pair of chords; for example, for the chords C major and A major, one of several possible $\boldsymbol{q}$ is $\boldsymbol{q} = (-1, 1)^{\intercal}$ because you can get from pitch class C to pitch class A by going down one perfect fifth and up one major third. An alternative is $\boldsymbol{q} = (3, 0)^{\intercal}$, while yet another is $\boldsymbol{q} = (7, 2)^{\intercal}$, and so forth (the $^{\intercal}$ symbol denotes the transpose operator that turns a row vector into a column vector, and vice versa, hence the just-mentioned vectors are column vectors);
- $\mathcal{Q}$ is the set (of infinite cardinality) of all such enharmonically equivalent $\boldsymbol{q}$ for any given $\boldsymbol{x}$ and $\boldsymbol{y}$;
- $\boldsymbol{c} = (\frac{1}{3}, -\frac{2}{3})^{\intercal}$ is the spatial displacement between between major and minor triads with the same root in a square *Tonnetz*;
- $g \in \{-1, 0, 1\}$ determines when the displacement vector $\boldsymbol{c}$ needs to be used, and in which direction: $g = 0$ is used when chords $\boldsymbol{x}$ and $\boldsymbol{y}$ are both major or both minor, $g = 1$ is used when chord $\boldsymbol{x}$ is major and chord $\boldsymbol{y}$ is minor, $g = -1$ is used when chord $\boldsymbol{x}$ is minor and chord $\boldsymbol{y}$ is major.

To more simply explain the equation, let us initially focus on a single possible value of $\boldsymbol{q}$. The term $\boldsymbol{q} + g\boldsymbol{c}$ corresponds to a displacement, in a square *Tonnetz*, between $\boldsymbol{x}$ and $\boldsymbol{y}$ after taking into account whether they are major or minor triads. The multiplication by $\boldsymbol{L}$ transforms this displacement into what would occur in the canonical hexagonal lattice.

linear map from 5-limit just intonation to 12-TET (Milne, Sethares, and Plamondon 2008).

The multiplication by **HS** first scales, then shears this displacement as a function of the two free parameters h and s. The Euclidean norm $\|\cdot\|_2$ of the resulting displacement is then used to provide its length – this gives the distance between the two chords in the sheared and scaled *Tonnetz*. This is done for all possible different values of $\boldsymbol{q}$ in $\mathcal{Q}$ (in the computational routine a finite, but sufficiently comprehensive, subset of $\mathcal{Q}$ is actually searched). From these, the minimum possible distance value is chosen, which is $d_{\mathrm{T}}(\boldsymbol{x}, \boldsymbol{y}; h, s)$.

In the approach described above, we model perceived chordal distances with their Euclidean distances in the *Tonnetz* (those distances varying as a function of the two parameters). This is a familiar tradition within psychology and perceptual research where empirical data are represented in a low-dimensional structure that ensures Euclidean distances correspond – as much as possible – with perceived distances (Shepard 1987). Such procedures have formed an important part of music perception research and computational visualization of musical structure (e.g. Shepard 1982; Krumhansl and Kessler 1982; Krumhansl 1998; Toiviainen and Krumhansl 2003). Given the simplicity and intuitive appeal of Euclidean distance, it is likely this underlies many historical representations too.

The neo-Riemannian approach (e.g. Hyer 1995; Cohn 1998; Tymoczko 2012) is somewhat different; here, the *Tonnetz* is often thought of as an approximate visualization of four important harmonic relationships – the parallel (e.g. C maj–C min), *leittonwechsel* (e.g. C maj–E min), relative (e.g. C maj–A min), and dominant (e.g. C maj–G maj) – all of which are spatially close in the hexagonal *Tonnetz*. This underlying "transformational distance" can be modelled simply by counting the minimum number of such transformations to get from chord to another (as in Krumhansl 1998, where such a model successfully predicted interkey distances derived from probe tone data). Clearly, this count is invariant with respect to the shear and scaling parameters we apply in the previously described *Tonnetz* model. We could parametrize the transformational model by allowing each transformation type to have a separate weight, but this would make it excessively flexible (over-parametrized) for our data set. We refer to this (unparameterized) model as the *transformational distance*.

2.2. *Voice-leading distance and its parameterization*

We utilize two types of voice-leading model that we term the *standard voice-leading model* (detailed in Sec. 2.2.1) and the *minimal voice-leading model* (detailed in Sec. 2.2.2). The standard model uses the actual pitch distances moved by each voice in the stimuli, while the minimal version uses the smoothest possible voice-leading (achieved by octave transposition and permutation of the voices) and so abstracts over the specific voicings actually used. At first sight, the standard voice-leading model seems more obvious and more likely to be useful as a representation of perceived triadic distance. However, the more abstract minimal version reflects a psychologically plausible prototype-based model of memory retrieval. Under such a model, any pair of chords with a specific voicing is considered to be treated cognitively as an elaboration, or exemplar, of a prototypical pair of chords, which themselves have a minimal voice-leading distance. In other words, we might think of each concretely heard chord pair as triggering a memory of a revoiced prototypical pair of chords, where the prototypical pair is such that voice-leading is minimized. Under such a model, it is the minimized distance that is reported by listeners. It is possible that cognitive processes reporting these two models of voice-leading distance (concrete vs abstract) may both be present simultaneously, an idea

which can be replicated by some linear combination of the standard and minimal voice-leading models. For these reasons, and because minimal voice-leading models are the type most widely discussed in contemporary music theory (e.g. Tymoczko 2011), both types of voice-leading model are here considered.

A principal difference between the predictions made by the two models is that the voice-leading distances between different voicings of the same underlying chords will typically be different for the standard version, whereas they will be identical for the minimal version. This means that for the experimental stimuli it is useful to include voicings of given chord pairs that do not always minimize the standard voice-leading distance; indeed, minimizing standard voice-leading distance is impossible if we also wish to satisfy other musical constraints such as avoiding parallel fifths. Furthermore, although minimal voice-leadings are common in real-world music, they are far from ubiquitous.

When given any two chords, each model produces a voice-leading vector, which needs to be converted to a distance value. Recall that in order to test all five models in their best possible light, we have striven to account for all common variations in the models when calculating distances. Consequently, rather than use a single fixed distance metric to calculate distances in the the voice-leading models, we use the very general and parameterizable p-norm, which includes Manhattan and Euclidean distance as two special cases (the use of different p-norms for voice-leading distances has been previously discussed by Callender 2004 and Tymoczko 2006, supporting online material). The two voice-leading models are now explained separately and in detail.

2.2.1. Standard voice-leading model

The pitch values in each chord are placed into a *pitch vector* in order of their voice which, for our stimuli, is cello, then viola, then second violin, then first violin. For convenience, we use MIDI pitch values, where C4 (middle C) has the value 60, and the units are twelve-tone equal temperament semitones (so D♭4 is 61). The *voice-leading vector* $\boldsymbol{v}$ is simply calculated by subtracting the pitch vector of the first chord $\boldsymbol{x}$ from that of the second chord $\boldsymbol{y}$ (i.e. $\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{x}$). For example, given the chord progression (C maj, E maj), as played by the specific voicings $\boldsymbol{x} = (48, 60, 64, 67)$ and $\boldsymbol{y} = (52, 59, 64, 68)$, the voice-leading vector is $\boldsymbol{v} = (4, -1, 0, 1)$.

The resulting voice-leading distance is calculated with a p-norm of this vector, where p is a free parameter that is optimized to the data. In other words, rather than a priori assuming which distance metric to use (represented by different values of p), we will use iteration to find the value of p that best fits the empirical data.

$$\begin{aligned} d_{\mathrm{sV}}(\boldsymbol{x}, \boldsymbol{y}; p) &= \|\boldsymbol{v}\|_p \\ &= \left(\sum_{n=1}^{N} |y_n - x_n|^p \right)^{1/p} , \end{aligned} \tag{2}$$

where $|\cdot|$ denotes the absolute value, and $p \in [1, \infty)$. For instance, when $p = 1$, the resulting distance is the Manhattan (or taxicab); when $p = 2$, the resulting distance is the Euclidean; when $p \to \infty$, the resulting distance is equivalent to the maximum distance moved by any voice. As p gets larger, larger elements of the voice-leading vector become progressively more important than smaller elements. At the limit $p \to \infty$, only the largest element matters. It is worth noting that when $p < 1$, the resulting value breaks the triangle inequality and so the function does not constitute a true norm nor accord

with most people's intuitive understanding of a reasonable distance measure.[5] Tymoczko (2006) points out various reasons why any reasonable model of voice-leading distance should support the triangle inequality. For this reason, we constrain this parameter to be no less than 1.

2.2.2. Minimal voice-leading model

For the minimal voice-leading model, the voice-leading distance between any two pitch vectors is given by choosing the minimum p-norm (for a given p) over the voice-leading vectors between all possible permutations of voices and octave transpositions of individual pitches in one of the chords. For example, the minimal voice-leading vector between $(\mathrm{C4}, \mathrm{E4}, \mathrm{G4}) = (60, 64, 67)$ and $(\mathrm{E4}, \mathrm{G5}, \mathrm{B5}) = (64, 79, 83)$ (an unlikely voice-leading, but useful for demonstration purposes) is not the actual voice-leading $(4, 15, 16)$, but the minimal $(1, 0, 0)$. This is achieved by permuting the voices of the second chord to $(\mathrm{B5}, \mathrm{E4}, \mathrm{G5}) = (83, 64, 79)$, and octave transposing two of the pitches to make $(\mathrm{B3}, \mathrm{E4}, \mathrm{G4}) = (59, 64, 67)$. There is no other such transformation of the second pitch vector that can minimize the voice-leading distance further. For the two N-element pitch vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, this can be expressed more formally as

$$d_{\mathrm{mV}}(\boldsymbol{x}, \boldsymbol{y}; p) = \min_{\substack{\boldsymbol{k} \in \mathbb{Z}^N \\ s \in \mathcal{S}_N}} \left(\sum_{n=1}^{N} |y_{s(n)} + 12k_n - x_n|^p \right)^{1/p}, \tag{3}$$

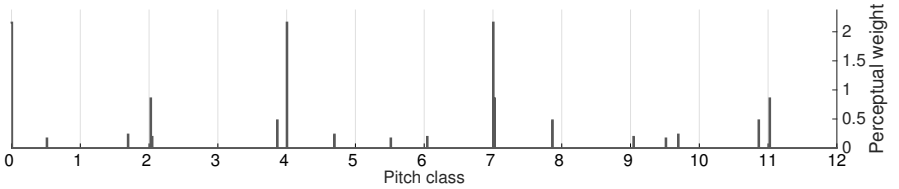
where $\boldsymbol{k} = (k_1, k_2, \ldots, k_N) \in \mathbb{Z}^N$ allows all possible octave transpositions of every pitch in $\boldsymbol{y}$ to be tested, s indexes over all $N!$ permutations of the set $\{1, 2, \ldots N\}$, and the set of all such permutations is denoted $\mathcal{S}_N$. For any given $p > 1$, the same transformation of $\boldsymbol{y}$ (i.e. permutation s and vector $\boldsymbol{k}$), will produce the minimal value for the resulting p-norm. This means the values of s and $\boldsymbol{k}$ that minimize the distance need to be determined for just one p-value for each pair of pitch vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. The vector $(y_{s(1)} + 12k_1 - x_1, y_{s(2)} + 12k_2 - x_2, \ldots, y_{s(N)} + 12k_N - x_N)$ that satisfies the minimality constraint is denoted the *minimal voice-leading vector*.[6]

Clearly, the minimal voice-leading may represent a useful abstraction across a broad equivalence class of chord voicings. In our model we apply one further abstraction, which is to convert our four-voice major and minor triads (in which one pitch class was played by two voices) into three-voice triads containing all three distinct pitches. For example, the actual chords $(48, 60, 64, 67)$ and $(41, 60, 65, 69)$ were represented by the vectors $\boldsymbol{x} = (0, 4, 7)$ and $\boldsymbol{y} = (5, 9, 0)$.

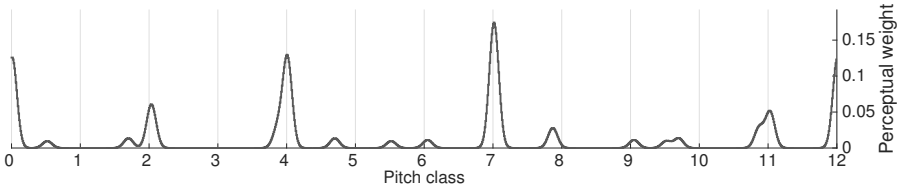
The resulting minimal voice-leading vectors are exactly the same as the *maximally smooth* voice-leadings calculated by Tymoczko's software application, called *Voice Leading*, which is available at http://dmitri.tymoczko.com/software.html. A maximally smooth voice-leading vector is one whose p-norm is minimal when $p = 1$ (i.e. the sum of its absolute-valued components is minimal). For certain chord pairs, there is more than one maximally smooth vector. Our voice-leading calculation utilizes the vector, from this set, that has the smallest range of values because this will additionally have the smallest

[5] The triangle inequality requires that $d(x, z) \leq d(x, y) + d(y, z)$; that is, the distance between two objects is at least as short as the distance when passing through a third object.

[6] Due to the abstraction (i.e. permutations) over voices, the ordering of elements in this vector has no meaning; hence this vector is most usefully thought of as just one out of all possible orderings of a single (unordered) multiset – the latter corresponding to the mathematical formalization used by Tymoczko (2006, 2011). Vectors may, however, be a more appropriate formalization whenever voices are not abstracted over.



(a) Unsmoothed spectrum.



(b) Smoothed spectrum.

Figure 2. Spectral pitch-class vectors showing the effect of smearing (convolving) a set of harmonic partials (from a major chord whose root is at pitch class 0) with a discrete normal distribution with a standard deviation of $\sigma = 6.83$ cents. The roll-off is $\rho = 0.75$. These are the parameter values as optimized to the experimental data, as detailed later.

p-norm when $p > 1$. For example, for the chord pair (C maj, D maj), Tymoczko's *Voice Leading* application outputs two maximally smooth voice-leadings: $(2, 2, 2)$ and $(2, 1, 3)$. We use the former because it has a smaller range of values and hence a lower p-norm for all $p > 1$. For example, the Euclidean norms (i.e. $p = 2$) of these two voice-leadings are 3.46 and 3.74, respectively (their norms are identical only when $p = 1$).

2.3. *Spectral pitch-class distance and its parameterization*

Spectral pitch-class distance utilizes the expectation tensors introduced in Milne et al. (2011). Here, the tensor is of the simplest kind—a *spectral pitch-class vector* in which delta spikes, indicating the log-frequencies modulo the octave and perceptual weights of all partials, are *smoothed* with a discrete normal distribution. This is illustrated in Figure 2.

The width of the smoothing is a free parameter σ, and the steepness of the roll-off in the weighting of ascending harmonics is another free parameter ρ. The smoothing-width parameter models the perceptual inaccuracies that result in close, but non-identical, frequencies being judged as having the same pitch class – the greater the width of the normal distribution the greater the modelled perceptual inaccuracy. The roll-off parameter models the lesser perceptual importance of higher partials relative to lower partials. This will likely depend on the spectrum used for the stimulus, but this parameter additionally allows the model to take account of psychoacoustic processes. For example, it is easier to perceptually resolve (consciously hear out) lower harmonics than it is higher harmonics, even when they have equal intensity (Bernstein and Oxenham 2003; Moore 2005).

More formally, for any given tone, a 1200-element row-vector of zeros is created. The first element represents the log-frequency (modulo the octave) of the pitch class of C. The second element is one cent higher, the third element is two cents higher, and so forth. This implies the vector encompasses a full octave range of finely-grained (cent-valued) pitch classes. For each of the tone's harmonics (indexed by i, so the fundamental has $i = 1$, the second harmonic an octave higher has $i = 2$, the third harmonic has

$i = 3$, etc.), a value of $1/i^\rho$ is added to the element corresponding to its cents value. For example, the first, second, fourth, eighth, and so on, harmonics of a complex harmonic tone all have the same pitch class value, so this element of the vector is given a total *weight* of $1 + 1/2^\rho + 1/4^\rho + 1/8^\rho + \cdots$. When $\rho > 0$, this means every higher partial contributes a lesser individual weight to the total than every lower partial, but no partial has a negative weight. The steepness of the roll-off is determined by the size of ρ. The same process is applied to all three tones in a chord, and the three resulting vectors of weights are summed to produce the row-vector $\boldsymbol{x}_{\text{wt}}$. An example of a typical such vector is illustrated in Figure 2(a).

The *spectral pitch-class vector* $\boldsymbol{x}_{\text{e}}$ is given by smearing $\boldsymbol{x}_{\text{wt}}$ across the log-frequency domain. This is achieved by circularly convolving with a discrete truncated normal distribution $\boldsymbol{g}$ with a standard deviation of σ; that is, $\boldsymbol{x}_{\text{e}} = \boldsymbol{x}_{\text{wt}} * \boldsymbol{g}$. The effect of this smearing is illustrated in Figure 2(b). For this vector, we use the term *pitch*, rather than *log-frequency* or *cents*, because the smoothing and weights are modelling perceptual processes that have "transformed" the original acoustical stimulus.

In our analysis, we include only the first twelve partials in the spectral pitch-class vectors. This is because partials higher than this typically cannot be perceptually resolved (Bernstein and Oxenham 2003), and removing them from the model reduces the number of calculations required (the computational efficiency of the model becomes a concern under optimization to the data, particularly when cross-validating). We would expect the optimized value of ρ to approximately correspond to the loudnesses of the partials in the sonic stimuli actually used. The string sounds used in our experiment have a typical amplitude (pressure) roll-off of $1/i$ which, using the rough approximation provided by Steven's law, corresponds to a loudness roll-of of about $1/i^{0.6}$. Hence, we would expect the optimized ρ to have a value similar to 0.6. As discussed in Milne et al. (2011, App. A, Online Supplementary), the standard deviation of $\boldsymbol{g}$ models the just noticeable frequency difference, which is 3–13 cents between 125 and 6000 Hz (Moore 1973). We would, therefore, expect the optimized smoothing width to be within or close to this range of cents values.

The *spectral pitch-class distance* of any two chords is simply modelled as the *cosine distance* between their respective spectral pitch-class vectors. *Cosine distance* is unity minus the cosine of the angle between the two vectors. For vectors all of whose values are positive (as is the case for spectral pitch-class vectors), their cosine distance is always between zero (maximally similar) and unity (maximally distant). Thus, for two triads $\boldsymbol{x}$ and $\boldsymbol{y}$, their spectral pitch-class distance d_{S} can be calculated as follows, given a roll-off parameter of ρ and a smoothing width parameter of σ

$$d_{\text{S}}(\boldsymbol{x}, \boldsymbol{y}; \rho, \sigma) = 1 - \frac{\boldsymbol{x}_{\text{e}} \boldsymbol{y}_{\text{e}}^{\intercal}}{\sqrt{\boldsymbol{x}_{\text{e}} \boldsymbol{x}_{\text{e}}^{\intercal} \; \boldsymbol{y}_{\text{e}} \boldsymbol{y}_{\text{e}}^{\intercal}}}, \tag{4}$$

where both $\boldsymbol{x}_{\text{e}}$ and $\boldsymbol{y}_{\text{e}}$ are row vectors, and $^{\intercal}$ is the transpose operator that converts a row vector into a column vector (and vice versa).[7]

[7] It is worth noting that the model described here is quite different to a voice-leading type model – even if each harmonic were to be represented as an individual "voice." For discussion of the differences between the *category domain pitch embeddings* used in voice-leading models and the *pitch domain embeddings* used in expectation tensors, and their implications when used to obtain distances, see Milne et al. (2011).

2.4. *Hamming distance benchmark model*

The final model – the *Hamming distance* between chords – serves as a simple "benchmark." As outlined earlier and detailed below, it has the benefit that each of the other models can be interpreted as an extension of it and makes use of a greater set of information. If the hypothesis underyling each of the other models is relevant to the data at hand, it should predict these data better than the Hamming model alone.

The Hamming distance model simply counts the number of non-zero entries in the minimal three-voice voice-leading vector, as defined above (though, in this case, the voice-leading vector representing the smallest Hamming distance – the vector with the fewest non-zero entries – is chosen). For example, the Hamming distance between the three-voice chords C maj $= (60, 64, 67)$ and A min $= (60, 64, 69)$ is 1 because only one voice moves; the distance between C maj $= (60, 64, 67)$ and A maj $= (61, 64, 69)$ is 2 because two voices move; the Hamming distance between C maj $= (60, 64, 67)$ and D min $= (62, 65, 69)$ is 3 because there are no pitch classes in common. The Hamming distance model has no parameters other than the linear intercept and slope parameters, also used by every other model so far discussed, to linearly map their values to the numerical values of the rating scale (which ran from 1 to 5).

More formally, we can represent the Hamming distance with

$$d_{\mathrm{H}}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\substack{\boldsymbol{k}\in\mathbb{Z}^N \\ s\in\mathcal{S}_N}} \sum_{n=1}^{N} 1 - \delta\big(y_{s(n)} + 12k_n - x_n\big)\,, \tag{5}$$

where $\delta(\cdot)$ is the Kronecker delta function, which is 1 when its argument is zero, but is otherwise 0. The remaining notation is as described for Equation (3). The chords are all entered as three-note versions, as in the minimal voice-leading model.

2.4.1. All four models viewed in detail as extensions of Hamming model

As already noted, the Hamming model serves as a particularly useful benchmark because, in important respects, all the other models can be thought of as alternative elaborations or extensions of it. For example, like Hamming, the voice-leading models give zero distance to pitch classes that do not change but, instead of giving a unit weight to any moving voice, they quantify it by the distance that voice actually moves. In this sense, voice-leading refines the Hamming model by making use of a greater amount of information.

In the case of the spectral pitch-class distance model, given a roll-off approaching infinity ($\rho \to \infty$), the resulting spectral pitch-class vectors only contain the fundamental pitch class of each chord tone (the harmonics are ignored). This means the resulting model is equivalent to the Hamming model (it just counts differences between the two vectors). However, as the roll-off parameter is reduced towards unity, the influence of the harmonics increases and the two models diverge. With a finite value of ρ, the spectral pitch-class distance model incorporates information not included in the Hamming model – the harmonics of every chord tone. However, this additional information is quite different to that utilized in the voice-leading models.

In the canonical hexagonal *Tonnetz*, all major and minor triads with two common tones (as represented by C–c, C–e, c–E♭) have a Euclidean distance of 1 (the same as their Hamming distance).[8] Major and minor triads triads with one common tone have

[8]For concision here, and in future examples, we use uppercase for major triads and lowercase for minor triads.

a Euclidean distance of $\sqrt{3} \approx 1.73$ (C–E♭, C–E, C–F, c–e♭, c–e, c–f), or 2 (C-c♯, C–f, c–F). Major and minor triads with no common tones have a distance of $\sqrt{7} \approx 2.65$ (C–d, C–e♭, C–f♯, c–D♭, c–E), or 3 (C–D♭, C–D, c–d♭, c–d), or $2\sqrt{3} \approx 3.46$ (C–F♯, c–f♯), or $\sqrt{13} \approx 3.61$ (c–D). So, like Hamming, the *Tonnetz* chord distances fall into non-overlapping groups as characterized by their number of common tones, but there is some additional variation within the one-common-tone group and the no-common-tone group. This can be thought of as additional structure beyond the Hamming that results from representing chord relationships in a regular two-dimensional geometry. The correlation between the canonical hexagonal chord *Tonnetz* and the Hamming model is very high ($r(24) = .96$ and, with the small shear and scale values of 0.17 and 1.08 respectively, the correlation can be maximized to $r(24) = .97$). By choosing different values of shear and scale, the *Tonnetz*' additional structure changes in extent and form. In these ways, the chord *Tonnetz* model represents an elaboration of the Hamming model.

As detailed in the next section, our experiment elicited ratings of the perceived distance between different pairs of root-position major and minor triads. If any of the three main models were to perform no better than the Hamming model at predicting these data, this would demonstrate that the extra information it takes into account – *voice-leading distance*, *non-fundamental harmonics*, or *a regular structure founded on fifths and thirds* – is irrelevant to these data.

3. The experiment

3.1. *Participants*

There were 35 participants (19 male, 16 female, with a mean age of approximately 30 years), most of whom were international students or staff of Jyväskylä University, Finland. Participants were asked to rate their instrumental and music theory skills. The average level of both was "intermediate," and only two participants had no playing or music theory skills (on a scale of "none" = 0, "basic" = 1, "intermediate" = 2, "advanced" = 3, average instrumental skill was 2.3, average music theory skill was 2.1).

3.2. *Apparatus*

The experimental interface was created with Max/MSP. The music was stored as MIDI files and played through a software sampler to emulate a naturalistic stereo recording of a string quartet (see Sec. 3.3.3). The synthesizer was Cakewalk's Dimension Pro playing a sample set from Garritan, and each instrument was individually panned to a left-right location such as one would typically hear in a commercial recording. The music was played over closed-back headphones (Audio Technica ATH-M40fs) in a quiet room.

3.3. *Stimuli and procedure*

As outlined in the introduction, the experiment was designed specifically to probe the posited underlying symmetrical aspects of harmony perception described in that section. We now discuss these methods in detail, and point out how they differ from previous related research conducted by Krumhansl and Kessler (1982), Bharucha and Krumhansl (1983), Bigand, Parncutt, and Lerdahl (1996), Krumhansl (1998), and Rogers and Callender (2006).

3.3.1. Avoiding order asymmetries in rating

As previously discussed, although asymmetries clearly exist in music and are a vital aspect of tonal cognition, our tests focus on stimuli and conditions well suited to comparing the five models being tested here (all of which are symmetrical with respect to order and transposition/translation). Thus, when asking participants to carry out ratings, we used adjectives that do not imply an ordering, and which refer to a relationship between the two chords, not to an affect induced by just one of them. We do not, for example, ask "how well the second chord follows the first" (as in Bharucha and Krumhansl 1983), or the degree of "tension" in one chord out of a pair (as in Bigand, Parncutt, and Lerdahl 1996). Of course, the chords in a pair may have differing degrees of tension, but our goal is to focus on distance, not asymmetrical features. For each pair of chords, we asked our participants two questions: "how 'similar' or 'dissimilar' do the two chords sound?," and "how 'well' or 'badly' do the two chords fit together?" In both cases, the question refers to the relationship between the two chords, not to a property of a single chord in the pair.

Furthermore, to minimize unwanted order asymmetries in the experiment, we played each pair of chords in a continuous loop (chord1–chord2–chord1–chord2–...). This differs from the method used by Bharucha and Krumhansl (1983) and Rogers and Callender (2006), where each stimulus comprised two chords played in a unidirectional sequence. A question remains as to whether musical distance is inherently asymmetric – this question cannot be answered by the given experimental setting, because it minimizes asymmetry.

3.3.2. Avoiding tonal context

We endeavoured, as much as practicable, to present the chord pairs without any previously established tonal context. As previously discussed, such contexts can induce unwanted asymmetries with respect to order and transposition. For each chord pair, the overall pitch was randomly transposed (with a uniform distribution) over a one-octave range of equally tempered semitones. In between each stimulus, a three-second randomly generated four-part atonal progression of six chords was played. The purpose of this was to displace any sense of a specific tonal centre that may have been inadvertently established by the previous chord pair. The order of presentation was also random for each participant. This differs from the method of Bigand, Parncutt, and Lerdahl (1996), where a tonal context was deliberately established.

3.3.3. Naturalistic stimuli

The musical examples played to participants used standard (common-practice) voice-leadings, and they were played with high-quality samples of the four instruments in a string quartet (two violins, a viola, and cello). Each instrument was independently panned to sound like a commercial recording of a string quartet. These naturalistic qualities not only enhance ecological validity, they additionally have two important impacts.

Firstly, the stereo and vibrato independence of the instruments playing the constituent notes of the chords should help to enhance the ability to separately stream each voice (Bregman 1990). For the same purpose, we also follow conventional rules of voice-leading by, for example, retaining common tones and avoiding parallel fifths and octaves (the utility of common-practice voice-leading to individuate voices is extensively discussed in Huron 2001). This is essential to test the voice-leading models fairly, while in no way disadvantaging the other models being tested.

Secondly, the spectral pitch-class distance model posits that the entire harmonic spec-

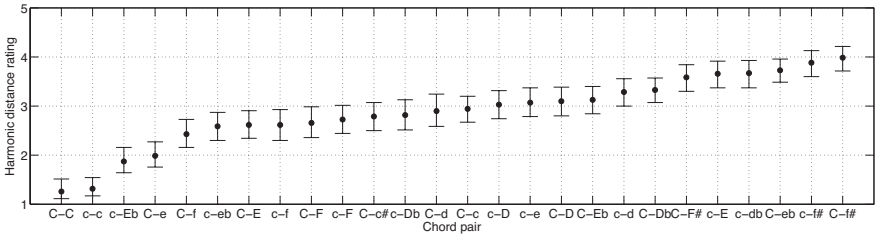


Figure 3. Mean distance ratings for all twenty-six chord pairs, and 95% confidence intervals.

trum associated with the played chords affects human judgement of triadic distance. If this should be the case, then naturalistic stimuli with harmonic complex tones are essential to test this model fairly, unlike, for example, the octave complex tones used in experiments such as Krumhansl and Kessler (1982), Bharucha and Krumhansl (1983), and Rogers and Callender (2006).[9] Stimuli with naturalistic harmonic complex tones in no way disadvantage the other models being tested, and were historically ubiquitous at the foundation of all of the other models.

All chords were played at a moderate tempo of 100 beats per minute (so the inter-onset intervals were 1200 ms). The note durations were 1125 ms, resulting in a "legato" articulation ratio of .94. Every chord was played in twelve-tone equal temperament.

The parts were composed so as to broadly follow conventional rules of voice-leading (given that both chords are always in root position): retention of common tones, choosing stepwise motion over leaps, avoiding parallel and hidden fifths, octaves and unisons, particularly in the bass and soprano parts. Due to the necessity for all chords to be in root-position, these rules cannot all be simultaneously fulfilled, so aesthetically based choices had to be made. There was no deliberate attempt made to minimize the standard voice-leading; rather, the concern was to produce musically reasonable sounding progressions. The MIDI files used are shown in score form in Figure 4 (note that for each presentation, each chord pair was randomly transposed over the range of −6 to +5 semitones).

3.3.4. Measurement

As discussed earlier, we probe the perceived distance of the chords in each stimulus by asking our participants two questions. One is a direct question about distance/dissimilarity. The other question, which is about *fit*, is commonly used in music perception research to measure the perceived distance between tones and/or chords in a variety of contexts (e.g. Krumhansl and Kessler 1982, Castellano, Bharucha, and Krumhansl 1984, Kessler, Hansen, and Shepard 1984, and the multidimensional scalings thereof). Our participants gave their responses on five-point scales with bipolar labels at the top and bottom – "dissimilar" and "similar," and "bad fit" and "good fit." We used the terms "similarity" and "dissimilarity" rather than "near" and "far" or "close" and "distant," because we did not want to inadvertently emphasize a one-, two-, or three-dimensional distance relationship due to the latter terms' conventional usage in spatial contexts.

In psychological measurement, it is common practice to use more than one question,

[9] A *harmonic complex tone* comprises frequency components at integer multiples of a fundamental frequency, an *octave complex tone* comprises only those frequency components at 2^n multiples of the fundamental – the first, second, fourth, eighth, sixteenth, etc., harmonics. Most pitched Western musical instruments, and the sung human voice, produce harmonic complex tones. Octave complex tones do not occur in nature and must be artificially synthesized.

and associated rating scale, and to combine them to produce a more complete measurement of a single underlying (latent) variable. Typically, it is advised that scales should be combined so long as they are sufficiently correlated (e.g. Cronbach's $\alpha > .7$), though it is also desirable that the two measures are not too highly correlated otherwise they may not be providing a sufficiently broad measurement (Iacobucci 2001). As discussed in the subsequent results section, participants' responses to the two questions were sufficiently – and not excessively – correlated, so we took their mean to serve as our final measure, and operationalization, of *perceived triadic distance*.

3.3.5. Minimizing confounds

In order to minimize the potential confound of chords with differing levels of harmonic consonance and dissonance, only root-position major and minor triads were used. In Western music theory, these are the only chords categorized as fully consonant – inversions of major and minor triads are generally described as unstable in effect, and other simultaneously sounded pitch-class sets, like diminished and augmented triads, or extensions like sevenths and ninths, are considered dissonant. Not having to account for consonance and dissonance makes the modelling simpler; furthermore, there is as yet no strong agreement about which are the most effective models of consonance and dissonance.

Ignoring overall transposition, the ordering of the chords, and their voicings, and assuming the enharmonic equivalences embodied in twelve-tone equal temperament, there are twenty-six different pairs of root-position major and minor triads. The following list shows these pairs notated such that the first chord is always C (C maj) or c (C min): C–C; C–D♭; C–D; C–E♭; C–E; C–F; C–F♯; c–c; c–d♭; c–d; c–e♭; c–e; c–f; c–f♯; C–c; C–c♯; C–d; C–e♭; C–e; C–f; C–f♯; c–D♭; c–D; c–E♭; c–E; c–F. When these labels are used in the figures below, they should be understood to refer not to the specific pair shown, but as representing an *equivalence class* containing all reasonable transpositions of both chords by the same amount (e.g. c–d ≡ c♯–d♯), over all enharmonic equivalences (e.g. C–e♭ ≡ C–d♯), and in both possible orderings (e.g. C–E ≡ E–C). Hence, the chord pairs C–E and C–A♭, for example, are in the same equivalence class because they are are related by a combination of transposition and reordering. The precise chords played to the participants, and the equivalence chord classes they represent are shown in Figure 4.

Each participant was presented with some randomly selected practice examples before starting. Each stimulus was then rated once for *similarity*, and once for *fit*. The experiment took approximately ten minutes, so any chance of listener fatigue was minimized.

4. Results

4.1. *Interparticipant correlations and preprocessing of data*

We calculated Cronbach's α to estimate the reliability of participants' responses for each of the two items separately (*similarity* and *fit*).[10] Across participants' ratings of "similarity," it was $\alpha = .97$; across ratings of "fit," it was $\alpha = .94$. In neither case, was there any participant whose removal increased the Cronbach's α (taken to two decimal

[10] Cronbach's α is equivalent to the mean split-half correlation of a data set and is used as an estimate of reliability, consistency, or homogeneity. A split-half correlation is given by splitting the data into two equally-sized halves, summing across each half, and then correlating these two summed halves. The mean split-half correlation is the mean correlation over all possible equal splits. It is equivalent to Cronbach's α, though the latter is calculated in a computationally simpler – though intuitively less understandable – manner.

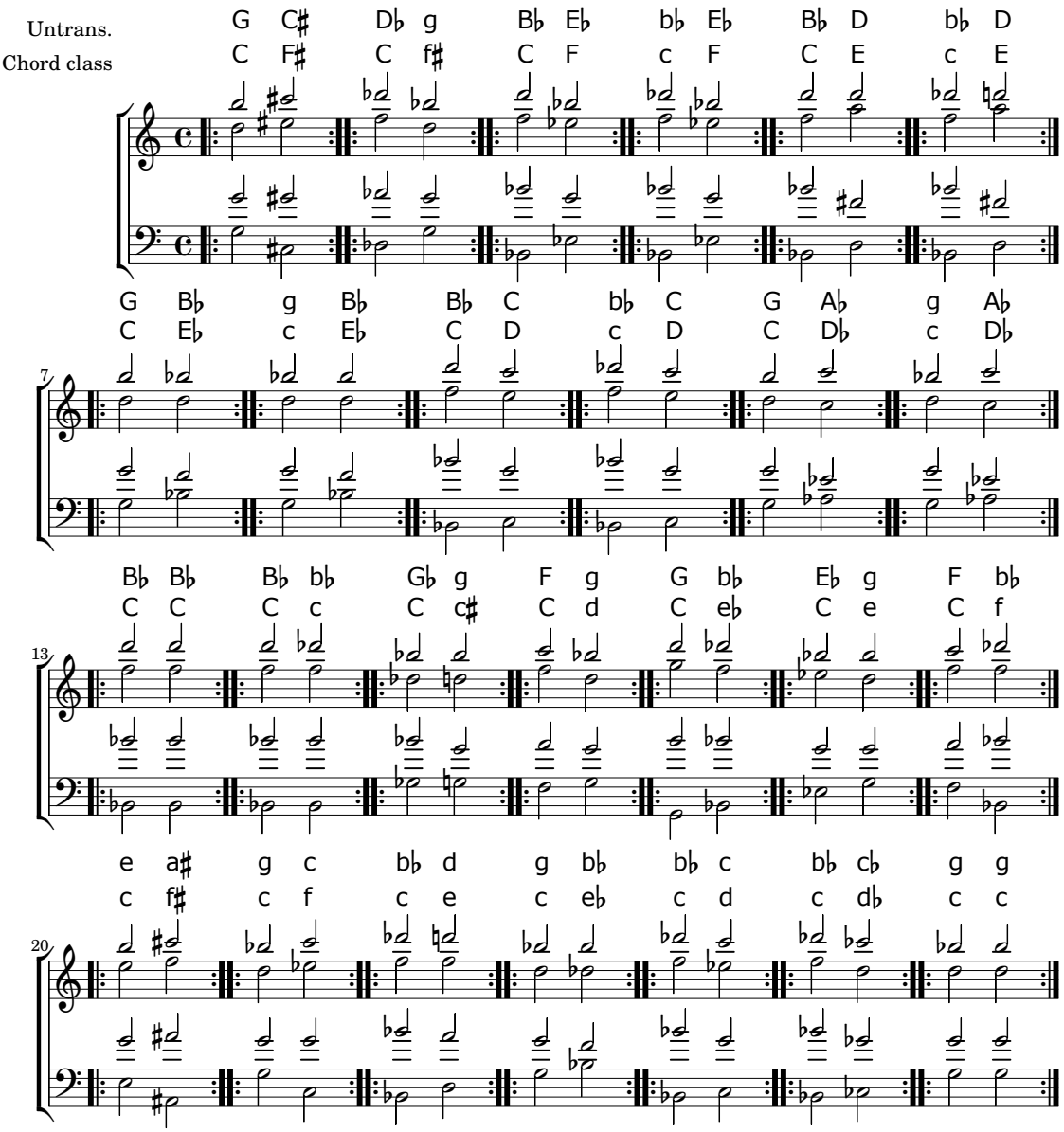


Figure 4. Prior to each chord pair being randomly transposed over the range −6 to +5 semitones, these are the 26 pairs used in the experiment. The chords' names corresponding to the notated pitches are shown on the top line above the staff; the corresponding chord equivalence classes' names are shown on the line below.

places). In light of these results, we took the means of all participants' responses for both fit and similarity to produce two mean rating scales.

Analysis of the two resulting scales demonstrated that *similarity* and *fit* are consistent with each other ($\alpha = .83$), consequently we averaged the two scales into a single scale called *triadic distance* (their high consistency was additionally indicated by the Cronbach's α of .97 across both similarity and fit, prior to taking their means).

In Figure 3, we show the resulting triadic distance ratings given to the twenty-six different chord pairs. They are arranged in order of their distance value (1 corresponds to "similar" and "good fit"; 5 corresponds to "dissimilar" and "bad fit"). Each value is surrounded by a 95% confidence interval calculated over 1000 bootstrap resamplings of

Table 1. Inter-correlations between the empirical data and all models (24 degrees of freedom).

	data	d_{H}	d_{sV}	d_{mV}	d_{Tr}	d_{S}	d_{T}
Experimental data	1.00	.88	.62	.72	.83	.91	.92
d_{H} – Hamming	.88	1.00	.70	.85	.82	.96	.89
d_{sV} – standard voice-leading	.62	.70	1.00	.62	.54	.63	.59
d_{mV} – minimal voice-leading	.72	.85	.62	1.00	.69	.79	.78
d_{Tr} – transformational	.83	.82	.54	.69	1.00	.93	.94
d_{S} – spectral pitch class	.91	.96	.63	.79	.93	1.00	.97
d_{T} – chord *Tonnetz*	.92	.89	.59	.78	.94	.97	1.00

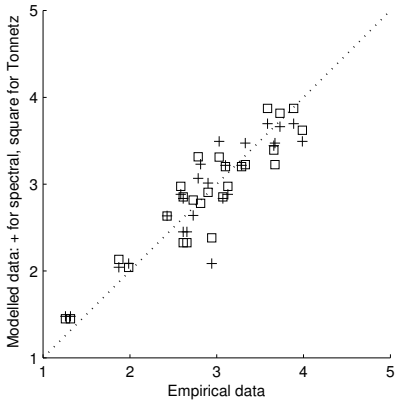


Figure 5. A scatter plot showing the relationship between the empirical data (x-axis) and their modelled values (y-axis). The spectral pitch-class model is indicated with the + symbol, the *Tonnetz* model with the □ symbol. A perfectly fitting model would have all data points on the diagonal dotted line. The vertical distance between each data point and this line is the model's error for that chord pair's perceived distance. For both models, the most outlying data point, with an empirical rating of almost 3, is the triad pair C–c. The root-mean-square-errors of the *Tonnetz* and spectral models are .28 and .30, respectively; their mean absolute errors are .22 and .23, respectively.

participants.[11]

4.2. *Model fitting and cross-validation*

We fitted all six models to the above triadic distance data using iterative optimization (MATLAB's `optimset` function) to minimize the sum of squared errors between the model and data. This maximizes the Pearson correlation between the model and the data and is equivalent to the fitting criterion used in standard linear regression. The resulting correlations, between all models and data, are shown in Table 1.

Of the voice-leading models, the minimal performs better than the standard; furthermore, when both voice-leading models are included in a linear regression, the coefficient for the standard voice-leading model is small and insignificant. Of the *Tonnetz*-based models, the *Tonnetz* performs better than the transformational distance. For the sake of brevity in the following analyses, we consider only the better of each pair. This leaves the following models: Hamming, minimal voice-leading, spectral pitch class, and *Tonnetz*. A scatter plot comparing the *Tonnetz* and spectral pitch-class distance models' predictions with the twenty-six empirical data points is shown in Figure 5.

The optimized parameter values are shown in Table 2. In Figure 6, we show the chord *Tonnetz* that results from these optimized parameter values (the corresponding pitch-

[11] Bootstrapping is a method for estimating the variance of a statistical estimate – in this case the variances of the means. It makes no assumptions about the underlying distribution of the data.

Table 2. Parameter values as optimized to the empirical data.

	p-norm	roll-off ρ	smoothing σ	scale s	shear h
d_{mV}	1.00	-	-	-	-
d_{S}	-	0.75	6.83	-	-
d_{T}	-	-	-	0.55	−0.15

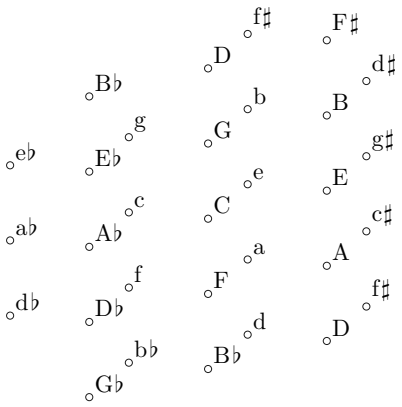


Figure 6. The optimized chord *Tonnetz* – major triads in uppercase, minor triads in lowercase.

class *Tonnetz* can be visualized by ignoring the minor chords and treating the major chord roots as pitch classes). In Figures 7 and 8, we show the spectral pitch-class distances for intervals and chords as generated under that model with optimized parameter values.

The optimized pitch-class *Tonnetz* predicts that perfect fifths are almost twice as close (perceptually) as major and minor thirds, and that major thirds are slightly closer than minor thirds. This is similar to conventional judgements of their relative consonances. Due to the construction of the *Tonnetz*, the distances of all other intervals are linear combinations of these fifths and thirds.

For the spectral pitch-class model, the optimized smoothing width falls within the expected range (3–13 cents), as does the roll-off, which corresponds approximately to the loudnesses of the partials in the string sounds used (similar optimized values were also obtained in the related models detailed in Milne, Laney, and Sharp 2015 and Milne, Laney, and Sharp 2016). As shown in Figure 7, the optimized spectral pitch-class distance model can calculate values for any interval size, including microtonal. In this context, we are interested in those that are twelve-tone equal temperament intervals and, hence, fall on the dotted radial lines (clock-face positions). For such intervals, the model predicts that, other than the unison/octave, only the perfect fifth/perfect fourth departs more than modestly from the maximal distance. This reflects the unique importance of this interval in Western music, and some non-Western music (Chalmers 1990; Xenakis 1992; Serrà et al. 2011).

Clearly, the models have differing flexibilities due to their differing constructions and nonlinear parameterizations. It might be that some of the models are achieving a high fit because they are excessively flexible and are fitting the noise in the data rather than the underlying process. A well-established method for determining this is to conduct multiple runs of k-fold cross-validation, and it is common to use a value of k approximating ten because this provides an effective trade-off between the bias and variance of the cross-validation estimates (e.g. Rodríguez, Pérez, and Lozano 2010).

We used 100 runs of 13-fold cross-validation, which means the empirical data set of 26

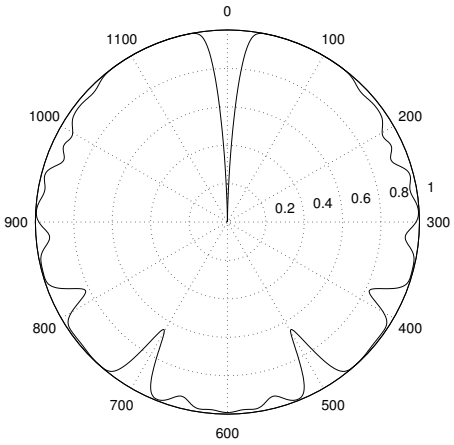


Figure 7. Spectral pitch-class distances of all pitch-class intervals. The size of the interval increases around the circle and is labelled in cents (one hundredths of a 12-TET semitone). The spectral pitch-class distance is 0 at the centre of the circle and 1 at the perimeter. The roll-off and smoothing parameters are as optimized to the data. The symmetry about the vertical axis results from the use of spectral pitch classes rather than pitches.

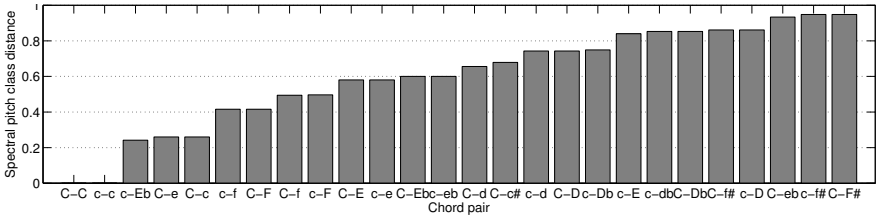


Figure 8. Spectral pitch-class distances, in order, of all chord pairs. The roll-off and smoothing parameters are as optimized to the data.

distance values is split into a *training set* of 24 distance values and a *validation set* of 2 distance values. The parameters of each model are optimized to the training set (for the Hamming model these are just the linear intercept and slope parameters; for the other models there are additional nonlinear parameters, as shown in Table 2). The modelled values for the two validation data points are then calculated. This procedure is done 13 times; in each case a different training and validation set is used, such that each validation set never contains a data point used in a previous validation set. This ensures we end up with 26 modelled values corresponding to all 26 data points. The cross-validation statistic of interest is then calculated for these values (in this case, the cross-validation correlation). Cross-validation statistics have an unknown variance, but this variance can be reduced by repeating the process multiple times with different validation sets and taking the mean value of the statistic. As mentioned above, we performed 100 runs of the 13-fold cross-validation.

The resulting cross-validation correlations are shown in Table 3. They replicate the rankings of the un-cross-validated results, and the top-performing models show only small reductions in their correlations after cross-validation. The Hamming model is a simple linear model having only intercept and slope parameters so we know, a priori, that it is not excessively flexible. The reductions in correlations (after cross-validation) of the spectral and *Tonnetz* models are no greater than that of the Hamming. This implies all three models (Hamming, spectral pitch-class distance, and *Tonnetz*) are not excessively flexible. In other words, they are not rendered vacuous by their parameterizations, and

Table 3. Cross-validation correlations between each model and the empirical data.

d_{H}	d_{mV}	d_{S}	d_{T}
.85	.66	.88	.90

Table 4. Partial F-tests to compare models.

	$\widehat{d_{\mathrm{mV}}}$	$\widehat{d_{\mathrm{S}}}$	$\widehat{d_{\mathrm{T}}}$
d_{H}	.665	.006*	.0003*
$\widehat{d_{\mathrm{S}}}$	-	-	.045

The row labels show the model included in the reduced regression, the column labels show the additional model included in the full regression. The entries show the p-value for the resulting partial F-test. Results that are significant at the .05 level after Bonferroni correction for multiple comparisons are starred.

are sufficiently parsimonious.

4.3. *Model comparisons*

In order to test the relative performances of the models, we fixed their nonlinear parameters to the previously optimized values (we denote these fixed models by $\widehat{d_{\mathrm{mV}}}$, $\widehat{d_{\mathrm{S}}}$, and so forth – note that the d_{H} has no nonlinear parameters). These predictors were then entered into a pairs of multiple linear regressions to enable partial F-tests to be performed. For example, to determine if the previously optimized *Tonnetz* model adds a significant amount of additional predictive power to the Hamming model, we compare a *full* regression model ($\mathit{distance} = \beta_0 + \beta_1 d_{\mathrm{H}} + \beta_2 \widehat{d_{\mathrm{T}}} + e$) and a *reduced* regression model ($\mathit{distance} = \beta_0 + \beta_1 d_{\mathrm{H}} + e$). A partial F-test allows us to determine whether the R^2 fit of the full model is significantly greater than that in the reduced model, given that the full model has additional degrees of freedom (in this case, there is one additional degree of freedom because the full model has an additional parameter β_2).

We are particularly interested in how voice-leading, spectral pitch-class distance, and the *Tonnetz* compare with the benchmark model. Given the results, we are additionally interested to know whether the *Tonnetz* model has significant predictive power beyond the spectral model. Table 4 summarizes the p-values for these four comparisons; all the F-tests have the degrees of freedom $F(1, 23)$.

In an investigation where multiple comparisons are made, the probability of a null hypothesis being incorrectly rejected increases. This can be corrected by dividing the significance level by the number of comparisons (this is the well-known Bonferroni correction). In our case, this means a p-value of $.05/4 = .0125$ is required for significance at the conventional 5% level. Under the Bonferroni correction, both the spectral pitch-class distance model and the *Tonnetz* model add a significant amount of additional predictive power beyond that provided by the Hamming benchmark. The minimal voice-leading model adds no significant predictive power beyond the Hamming (indeed, a partial F-test comparing a voice-leading-only model with a model also including the Hamming predictor shows the former to perform significantly worse, $p = .00005$). There is some evidence, though inconclusive due to the Bonferroni correction, that the *Tonnetz* model

may add predictive power beyond that provided by spectral pitch classes.

We do not report results from a multiple linear regression of the empirical data on all models, or even on the best models, because they are too highly intercorrelated (multicollinear) for the estimates of their coefficients to be reliable. For example, a regression on d_{H}, $\widehat{d_{\mathrm{T}}}$, and $\widehat{d_{\mathrm{S}}}$ has variance inflation factors (VIFs) of 16.7, 54.5, and 21.1, respectively (any VIF greater than 5 is typically considered to make coefficient estimates excessively unreliable).

5. Discussion

Let us start this section with the important caveat that any generalizations we make here can only be reliably extended to musical listeners with a similar university-educated and predominantly Western background as our participants. This is not to say that these results are not more generally applicable, but our data should not be used to generalize beyond this. Furthermore, the confidence intervals shown in Figure 3 suggest that, ideally, this or a similar experiment should be replicated in order to make firmer conclusions about the models' relative and absolute strengths.

The results, as summarized in Tables 1 and 4, demonstrate that the number of non-common pitch classes in two chords – the Hamming model – is a very effective predictor of their perceived distance. They also show that the the sizes of the pitch or pitch-class intervals between pairs of chords – the modus operandi of voice-leading models – do not, in this case, play a useful predictive role. Indeed, and surprisingly, the voice-leading model performs significantly worse than the Hamming model (we suggest a possible reason later). In contrast, both spectral pitch-class distance and the *Tonnetz* model perform very effectively and also perform better than the Hamming model. Indeed, it is notable that the *Tonnetz* – a simple music theory representation of Western harmony that dates back to the end of the Baroque period – is so effective for listeners exposed to centuries' worth of later music.

The *Tonnetz* is often explained as being based on the consonance of perfect fifths and major and minor thirds (and their inversions) – these intervals form the *Tonnetz*' axes (Hyer 1995; Gollin 2011) (this explanation for the *Tonnetz* can be agnostic as to why these specific intervals have high consonance; e.g. whether consonance is psychoacoustical, cultural, or both). However, these harmonic consonance based justifications are not directly applicable to the context explored here, which is the extent to which successive (non-simultaneous) chords fit, rather than any quality of simultaneously played harmonic intervals. The predictive effectiveness of spectral pitch-class distance ($r(24) = .91$) and the *Tonnetz* ($r(24) = .92$) and their very high mutual correlation ($r(24) = .97$) suggest that spectral pitch-class distance provides a sensory (psychoacoustic) explanation for perceived triadic distances and for the *Tonnetz* (the latter being a music theory representation of these perceived musical distances).

An alternative – and non-acoustic – explanation for the pitch-class *Tonnetz* comes from Balzano (1980) (as illustrated in Fig. 5 of that paper and the corresponding chord *Tonnetz* in Fig. 1(b) of this paper), who uses it as a representation of how the cyclic group of order 12 (the octave) is the direct product of the cyclic subgroups of orders 3 and 4 (minor and major thirds, respectively). The resulting lattice has orthogonal axes of major and minor thirds and is somewhat similar to the canonical pitch-class *Tonnetz* (indeed, in our parameterization, it is given by $h = 0$ and $s = \sqrt{3}$). In the acoustical explanation, scaling and shearing balance the relative consonances of perfect fifths and major and minor thirds. In the group theory explanation, the scaling and shear parameters have no

obvious justification because, in this purely mathematical context, all integers should be treated equally and the two subgroups – major and minor thirds – should be orthogonal. It is interesting to note that when the chord *Tonnetz* model's parameters are fixed to the above-mentioned values for the Balzano version, its correlation with the distance data drops from $r(24) = .92$ to only $r(24) = .69$, which is worse than the benchmark model and even the minimal voice-leading model. This suggests that the effectiveness of the *Tonnetz* model results from its ability to model acoustical rather than group theory properties.

This research provides some evidence, though not conclusive, that the *Tonnetz* may have greater predictive power than the spectral pitch-class distance model. If this were the case, this might even suggest that the brain summarizes implicitly learned spectral similarities in a lower-dimensional form that is essentially isomorphic to the *Tonnetz* (in terms of hypothesized neural implementation this might be a spatial isomorphism, as suggested by Janata et al.'s (2002) findings, or a functional isomorphism). However, this is speculative – it may be that, under replication, different results will obtain (spectral pitch-class distance may perform better than the *Tonnetz* or there may be a more significant difference). Another thing to consider, when comparing the *Tonnetz* and spectral distance, is that the former has narrower applicability than the latter. The latter can easily be generalized to any possible scale tuning or spectral tuning (e.g. the non-harmonic partials produced by many percussion and non-Western instruments), as investigated in Milne, Laney, and Sharp (2016). It is not obvious how, or if, the *Tonnetz* could be extended to cover such non-harmonic timbres and non-standard tunings.

Before finishing, let us consider some possible reasons for the comparatively weak performances of the voice-leading models. Due to our desire to respect common practice voice-leading rules (such as avoiding parallel fifths and octaves) and to use only root-position triads, the chord voicings used for given chord pairs in the experimental stimuli (see Fig. 4) often did not minimize their standard voice-leading distances (the voice-leadings were not maximally smooth). Might this have prejudiced the voice-leading models? For example, consider the chord loop notated as G–A♭, which was played with the voicing (G3, G4, D5, B5)–(A♭3, E♭4, C5, C♭6). This has a voice-leading vector of $(1, -4, -2, 1)$ whose 1-norm is 8 semitones. If parallel fifths and octaves were used instead, this could be voiced as (G3, G4, D5, B5)–(A♭3, A♭4, E♭5, C♭6) whose voice-leading vector $(1, 1, 1, 1)$ has a 1-norm of just 4. Similarly, the chord loop notated as (E♭3, G4, E♭5, B♭5)–(G3, G4, D5, B♭5) has a voice-leading vector with a 1-norm of 5. If a non-root-position chord were used instead, this could be voiced as (E♭3, B♭3, G4, B♭4)–(D3, B♭3, G4, B♭4) whose voice-leading vector's 1-norm is only 1. As noted in Sections 2.2 and 3.3.3, even leaving aside the constraints on parallels and inversions, voicings were not always chosen to minimize the standard voice-leading distance, for the following reason. Using sounded voice-leadings that are not maximally smooth is advantageous because the distances given by the standard voice-leading model and the minimal voice-leading model will be less highly correlated, thereby allowing their relative performance to be more clearly disambiguated.

If voice-leading distance – as conventionally described – does play a meaningful role for these data, there are three plausible ways it can function: a) we perceive chord distances directly from the voicings actually used (as reflected by the standard model), (b) the distances result from maximally smooth versions of the actual voice-leadings (as reflected by the minimal model), (c) distances result from some combination of the two (as reflected by a linear combination of the standard and minimal models). However, individually neither model performs well (relative to the benchmark), and the regression using both the minimal and standard models leads to no significant improvement beyond

the minimal.

Another possibility may be that the *outer voices* (bass and soprano) are more salient than the *inner voices* (tenor and alto), as experimentally demonstrated by Huron (1989). This was checked with a regression containing both an adjusted standard voice-leading model that considered only the outer voices, and the original standard voice-leading model. The correlation achieved with the experimental data improved from $r(24) = .62$ to $r(24) = .72$. This matches the minimal voice-leading model, though with one extra parameter, but still falls well short of the Hamming, spectral and *Tonnetz* models.

As discussed in the introduction, it is possible our participants were unable to individuate all or some of the four voices in each chord, which would weaken the impact of voice-leading, but this does not immediately explain why the voice-leading model would be performing worse than the Hamming. Figure 7 may provide an explanation. Let us consider the pitch-class intervals of sizes 1 to 6 (semitone to tritone). Spectral pitch-class distance indicates that (for 12-TET tunings) we perceive the voice-leading-large perfect fourth/perfect fifth to be closer than all (non-zero) voice-leading-small intervals. Indeed, the correlation between voice-leading size and spectral pitch-class distance over intervals from size 1 to 6 inclusive is negative ($r = -.37$). This suggests that spectral distance, in this case, completely overwhelms the actual pitch distance moved.

Another reason for the disappointing performance of the voice-leading models may simply be due to the constrained nature of the stimuli – in a longer succession of chords or a more contrapuntal context where harmony does not move in such a block-like manner, it may be easier to individuate the voices. In such contexts, therefore, we may find that voice-leading plays a more important role. It is interesting to note that in the experiment conducted by Bigand, Parncutt, and Lerdahl (1996), which used a succession of three chords, a voice-leading model performed comparatively well. However, that experiment differed in two other important respects – a tonal context was established, and participants were asked to rate the tension of a single chord in the progression (rather than a distance or fit relationship between the chords). For the reasons outlined in this paragraph, it would be wise to reserve final judgement on the effectiveness of voice-leading models until a wider range of musical material and different rating scales are experimentally tested.

With the exception of the standard voice-leading model (which was the worst performing), all the models used pitch classes rather than pitches (no distinction was made between log-frequencies an octave apart). It is possible that pitch-class models, such as these, will perform less well when the data also includes responses to chords separated by large pitch distances (e.g. greater than one or two octaves). But, such chord progressions are relatively uncommon, and it seems that using pitch classes is beneficial for modelling the perceived distances of the types of chord progressions typically used in music. An interesting possibility would be to develop the spectral model to allow for the spectral pitch (not pitch-class) vectors to be discretely smeared (convolved) across octaves as well as across closely neighbouring log-frequencies. However, this would require at least one additional free parameter, so a larger set of empirical data would be required to test it.

In general terms, given the current state of knowledge, it is unclear the extent to which human judgements of similarity and fit are determined by low-level psychoacoustic mechanisms, as opposed to learned cultural factors. However, this experiment demonstrates the remarkable extent to which spectral pitch-class distance – a straightforward model of perceptual pitch uncertainty applied to physically present harmonics – is able to account accurately for human judgements, effectively as well as or better than any available model. In summary, therefore, we feel this research adds evidence in support of sensory (psychoacoustic) processes underlying the perceived structure of harmonic relationships

in tonal music.

6. Conclusion

We have investigated a variety of models of the perceived (symmetrical) distance between root-position major and minor triads, and tested them against ratings given by participants. We used a set of methods to minimize the possibility that the principle embodied by any single model may gain an unfair advantage – for example, by using naturalistic stimuli and using parameterizations to allow for appropriate model flexibility (as confirmed by cross-validation).

The results indicate that the number of common tones between chords (abstracted across voices and octaves) is a highly effective predictor of their perceived distance. They also indicate that the harmonics of the two chords (their spectral pitch distance) play an important additional role but that, in this context, the pitch distances moved by the musical voices play no additional role (whether or not those pitch differences are abstracted over voices and octaves). We also show that the *Tonnetz* has a predictive effectiveness that is similar to spectral pitch-class distance, and that the two models have an extremely high correlation. We suggest that spectral pitch-class distance provides a sensory explanation for perceived triadic distances and their music theory representation, the *Tonnetz*.

Acknowledgements

We would like to thank the two anonymous reviewers for helpful comments and suggestions. The first author would also like to thank Tuomas Eerola and Petri Toiviainen for supporting this project as part of a Master's degree for the MMT programme at the University of Jyväskylä (the models used in the resulting thesis, Milne 2009, were quite different to those described in this paper), as well as the students, staff, and other volunteers who participated in the experiment.

Supplemental online material

Supplemental online material for this article can be accessed at doi-provided-by-publisher and http://www.dynamictonality.com/harmonic_distance_files/.

Disclosure statement

The authors have no conflict of interest.

References

Bachem, A. 1950. "Tone height and tone chroma as two different pitch qualities." *Acta Psychologica* 7: 80–88.

Balzano, Gerald J. 1980. "The group-theoretic description of 12-fold and microtonal pitch systems." *Computer Music Journal* 4 (4): 66–84.

Bernstein, Joshua G., and Andrew J. Oxenham. 2003. "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?." *The Journal of the Acoustical Society of America* 113 (6): 3323–3334.

Bharucha, Jamshed Jay, and Carol L. Krumhansl. 1983. "The representation of harmonic structure in music: Hierarchies of stability as a function of context." *Cognition* 13: 63–102.

Bigand, Emmanuel, Richard Parncutt, and Fred Lerdahl. 1996. "The perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training." *Perception and Psychophysics* 58 (1): 125–141.

Bregman, Albert S. 1990. *Auditory Scene Analysis*. Cambridge, MA, USA: MIT Press.

Callender, Clifton. 2004. "Continuous transformations." *Music Theory Online* 10 (3).

Callender, Clifton, Ian Quinn, and Dmitri Tymoczko. 2008. "Generalized voice-leading spaces." *Science* 320 (5874): 346–348.

Capuzzo, Guy. 2014. "Neo-Riemannian theory and the analysis of pop-rock music." *Music Theory Spectrum* 26 (2): 177–200.

Castellano, Mary A., J. J. Bharucha, and Carol L. Krumhansl. 1984. "Tonal hierarchies in the music of North India." *Journal of Experimental Psychology: General* 113 (3): 394–412.

Chalmers, John. 1990. *Divisions of the Tetrachord*. Frog Peak Music.

Chew, Elaine. 2006. "Slicing it all ways: Mathematical models for tonal induction, approximation, and segmentation using the spiral array." *INFORMS Journal on Computing* 18 (3): 305.

Cohen, David E. 2002. "Notes, scales, and modes in the earlier Middle Ages." In *The Cambridge history of Western music theory,* edited by T. Christensen. Cambridge, UK: Cambridge University Press.

Cohn, Richard. 1997. "Neo-Riemannian operations, parsimonious trichords, and their "Tonnetz" representations." *Journal of Music Theory* 41 (1): 1.

Cohn, Richard. 1998. "Introduction to neo-Riemannian theory: A survey and a historical perspective." *Journal of Music Theory* 42 (2): 167–180.

Deutsch, Diana. 1982. "The processing of pitch combinations." In *The Psychology of Music,* edited by Diana Deutsch. Academic Press.

Douthett, Jack, and Peter Steinbach. 1998. "Parsimonious graphs: A study in parsimony, contextual transformations, and modes of limited transposition." *Journal of Music Theory* 42 (2): 241–263.

Drobisch, M. W. 1855. "Über musikalische Tonbestimmung und Temperatur." In *Abhandlungen der Koniglich sachsichen Gesellschaft der Wissenschaften zu Leipzig. Vierter Band: Abhandlungen der mathematisch-physischen Classe. Zweiter Band,* Leipzig: S. Hirzel.

Euler, L. 1739. *Tentamen novae theoriae musicae ex certissismis harmoniae principiis dilucide expositae*. Saint Petersburg: Saint Petersburg Academy.

Gollin, Edward. 2011. "From matrix to map: *Tonbestimmung*, the *Tonnetz*, and Riemann's combinatorial conception of interval." In *The Oxford Handbook of Neo-Riemannian Music Theories,* edited by Edward Gollin and Alexander Rehding, chap. 9, 271–293. Oxford University Press.

Holland, Simon. 1994. "Learning about harmony with Harmony Space: An overview." In *Music Education: An Artificial Intelligence Approach,* edited by M. Smith, A. Smaill, and G. Wiggins. Springer-Verlag.

Huron, David. 1989. "Voice denumerability in polyphonic music of homogeneous timbres." *Music Perception* 6 (4): 361–382.

Huron, David. 2001. "Tone and voice: A derivation of the rules of voice-leading from perceptual principles." *Music Perception* 19 (1): 1–64.

Hyer, B. 1995. "Reimag(in)ing Riemann." *Journal of Music Theory* 39 (1): 101–138.

Iacobucci, Dawn. 2001. "Measurement." *Journal of Consumer Psychology* 10 (1&2): 55–69.

Janata, Petr, Jeffrey L. Birk, John D. Van Horn, Marc Leman, Barbara Tillmann, and Jamshed J. Bharucha. 2002. "The cortical topography of tonal structures underlying Western music." *Science* 298 (5601): 2167–2170.

Kessler, Edward J., Christa Hansen, and Roger N. Shepard. 1984. "Tonal schemata in the perception of music in Bali and in the West." *Music Perception* 2 (2): 131–165.

Krumhansl, Carol L. 1990. *Cognitive Foundations of Musical Pitch*. Oxford: Oxford University Press.

Krumhansl, Carol L. 1998. "Perceived triad distance: Evidence supporting the psychological reality of neo-Riemannian transformations." *Journal of Music Theory* 42 (2): 265–281.

Krumhansl, Carol L., and Edward J. Kessler. 1982. "Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys." *Psychological Review* 89 (4): 334–368.

Longuet-Higgins, H. C. 1962. "Two letters to a musical friend." *The Music Review* 23: 244–248.

McLachlan, Neil M., David J. T. Marco, and Sarah J. Wilson. 2012. "Pitch enumeration: Failure to subitize in audition." *PLoS ONE* 7 (4): e33661.

Milne, Andrew J. 2009. "A psychoacoustic model of harmonic cadences." Master's thesis, University of Jyväskylä, Finland.

Milne, Andrew J., Robin Laney, and David B. Sharp. 2015. "A spectral pitch class model of the probe tone data and scalic tonality." *Music Perception* 32 (4): 364–393.

Milne, Andrew J., Robin Laney, and David B. Sharp. 2016. "Testing a spectral model of tonal affinity with microtonal melodies and inharmonic spectra." *Musicae Scientiae* Advance online publication. doi: 10.1177/1029864915622682.

Milne, Andrew J., William A. Sethares, Robin Laney, and David B. Sharp. 2011. "Modelling the similarity of pitch collections with expectation tensors." *Journal of Mathematics and Music* 5 (1): 1–20.

Milne, Andrew J., William A. Sethares, and James Plamondon. 2008. "Tuning continua and keyboard layouts." *Journal of Mathematics and Music* 2 (1): 1–19.

Moore, Brian C. J. 1973. "Frequency difference limens for short-duration tones." *Journal of the Acoustical Society of America* 54: 610–619.

Moore, Brian C. J. 2005. *Introduction to the Psychology of Hearing*. London: Macmillan.

Parncutt, Richard. 1988. "Revision of Terhardt's psychoacoustical model of the root(s) of a musical chord." *Music Perception* 6 (1): 65–94.

Révész, Geza. 1913. *Zur Grundlegung der Tonspsychologie*. Leipzig: Veit.

Rodríguez, Juan Diego, Aritz Pérez, and Jose Antonio Lozano. 2010. "Sensitivity analysis of k-fold cross validation in prediction error estimation." *IEEE Transactions On Pattern Analysis And Machine Intelligence* 32 (3): 569–575.

Rogers, Nancy, and Clifton Callender. 2006. "Judgments of distance between trichords." In *Proceedings of the 9th International Conference on Music Perception and Cognition,* edited by Mario Baroni, Anna Rita Addessi, Roberto Caterina, and Marco Costa, 1686–1691.

Serrà, Joan, Gopala K. Koduri, Marius Miron, and Xavier Serra. 2011. "Assessing the tuning of sung Indian classical music." In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR),* Miami, FL, USA, 263–268.

Shepard, Roger N. 1982. "Geometrical approximations to the structure of musical pitch." *Psychological Review* 89 (4): 305–333.

Shepard, Roger N. 1987. "Toward a universal law of generalization for psychological science." *Science* 237 (4820): 1317–1323.

Toiviainen, Petri, and Carol L. Krumhansl. 2003. "Measuring and modeling real-time responses to music: The dynamics of tonality induction." *Perception* 32 (6): 741–766.

Tymoczko, Dmitri. 2006. "The geometry of musical chords." *Science* 313 (5783): 72–74.

Tymoczko, Dmitri. 2011. *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*. Oxford University Press.

Tymoczko, Dmitri. 2012. "The generalized Tonnetz." *Journal of Music Theory* 56 (1): 1–52.

Xenakis, Iannis. 1992. *Formalized Music: Thought and Mathematics in Composition*. Revised ed. Pendragon Press.